

The Advantages of Machine Aided Co-reference Resolution for Research Cruise Metadata

Adam Shepherd • June 1 2017

Co-authors: Cyndy Chandler, Robert Ako, Douglas Fils, Danie Kinkade.



Oceanographic Research Cruise



Oceanographic Research Cruise: 2 Ideas of a Cruise

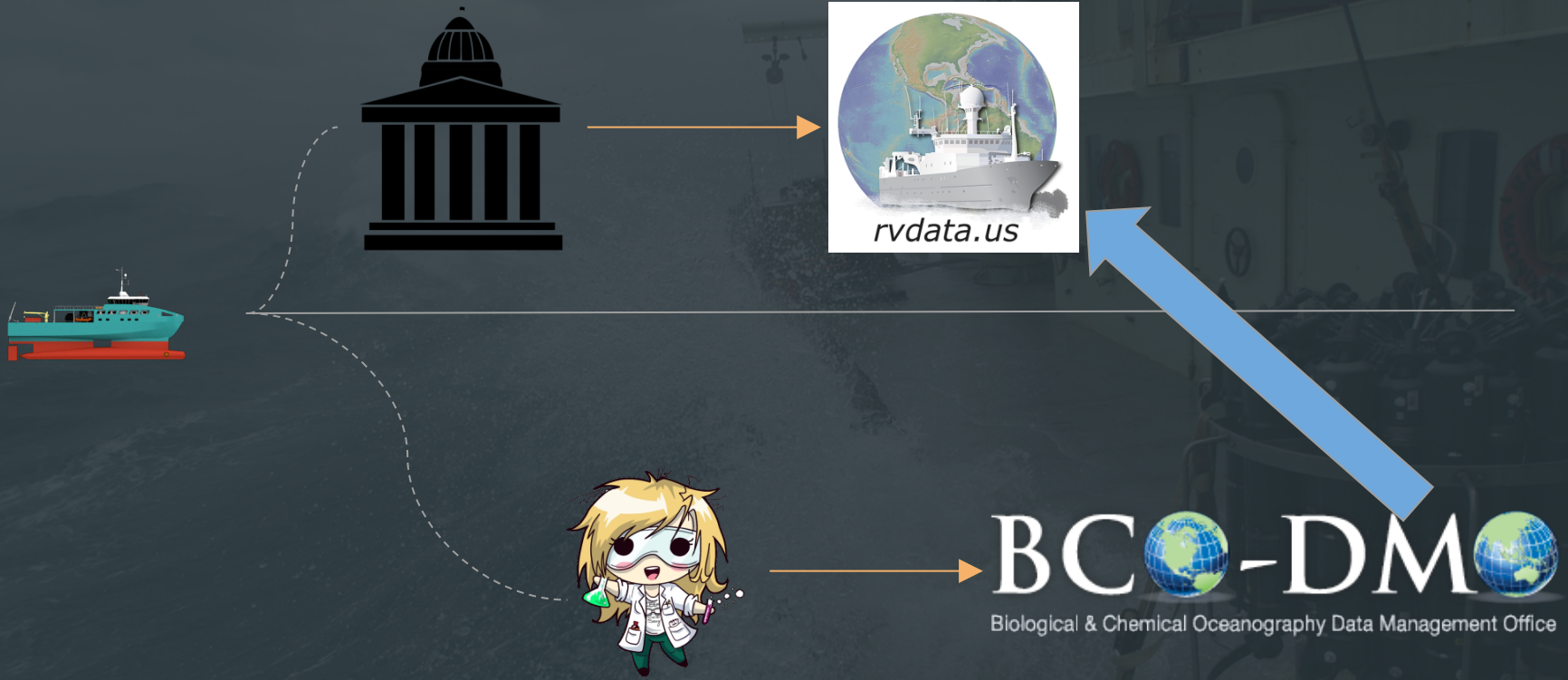


Oceanographic Research Cruise: Reporting Their Ideas



BCO-DMO
Biological & Chemical Oceanography Data Management Office

Oceanographic Research Cruise: Linking to Authorities



Oceanographic Research Cruise: Benefits of Linking

Deployment: AT11-07

Map It

Deployment: AT11-07

Chief Scientist: [Dr Hans Schouten](#) (Woods Hole Oceanographic Institution, WHOI)

Platform: [R/V Atlantis](#)

Platform Type: vessel

Start Date: 01/28/2004

End Date: 02/24/2004

Location: East Pacific Rise, Pacific

▾ [GeoLink](#)



Check GeoLink

▾ [Datasets](#)

Data from this Deployment only	Full Dataset (multiple deployments)
Assembled metagenome sequences > 500 bp (AT11-07)	Assembled metagenome sequences > 500 bp

Oceanographic Research Cruise: Benefits of Linking

Deployment: AT11-07

Map It

Deployment: AT11-07

Chief Scientist: [Dr Hans Schouten](#) (Woods Hole Oceanogra

Platform: [R/V Atlantis](#)

Platform Type: vessel

Start Date: 01/28/2004

End Date: 02/24/2004

Location: East Pacific Rise, Pacific

GeoLink



Check GeoLink

Datasets

Data from this Deployment only

Assembled metagenome sequences > 500 bp
(AT11-07)

Full Datas

Assemble
500 bp

Location: East Pacific Rise, Pacific

GeoLink



GeoLink Resource Information for: **AT11-07**

Cruise Information

Cruise: AT11-07

Identifier: [doi:10.7284/900008](#)

Start Port: San Diego

End Port: Puntarenas

Physical Samples (441)

Sample	Description	People	Dataset	Dates
1. IGSN:MGD000AVU	Elevation: -2512 Meters Feature of Interest: EPR:9N Feature Type: SpreadingCenterSegment Geometry: POINT(-104.291693 9.840713)	Edwards, Katrina Collector	AT11-07, Sampler>IncubationChamber	Collected: 2004-02-01 to Registered: 2012-03-11 Published: 2012-03-10
2. IGSN:MGD000B0A	Elevation: -2512 Meters Feature of Interest: EPR:9N Feature Type: SpreadingCenterSegment	Edwards, Katrina Collector	AT11-07, Sampler>IncubationChamber	Collected: 2004-02-01 to Registered: 2012-03-11 Published: 2012-03-10

Cruise DOI
(via R2R)

Hard rock geology
(via SESAR)

2,049

BCO-DMO Cruises.

6,555

R2R Cruises.

13,431,195

Comparisons.

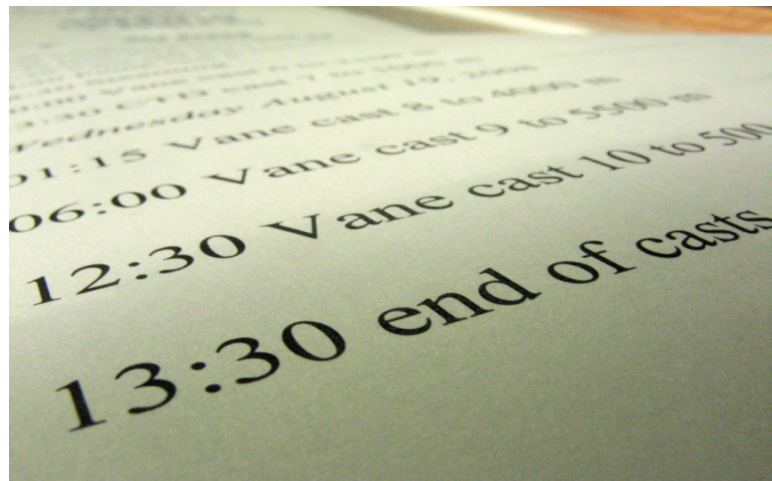
28

Years.

15 seconds each, 40 hrs/wk, 50 wks/yr

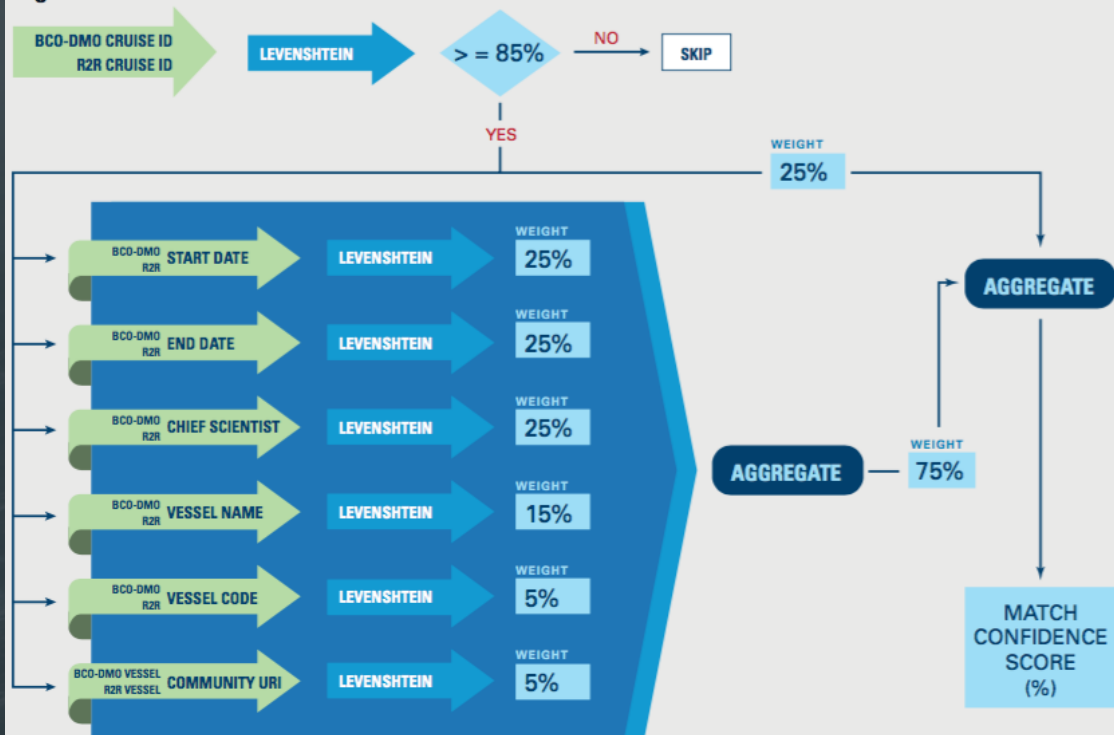
Entity Matching Ontology (EMO)

- Extends PROV-O
- Defines the properties of a 'Thing' that could be used for comparison
- Defines a potential match for a given Entity
- Hook for describing a matching algorithm



BCO-DMO / R2R EMO Extension

Algorithm



Check all BCO-DMO Cruises

CruiseID Match Threshold *

85

Set the floor on matching cruiseID scores. Valid values are between 1 and 100. The cruise ID matching score is a percentage where 100% represents a perfect match and 0% not matching at all

Calculate the combined cruise property match score

Weigh differences in cruise properties. The total of all weights should equal 100%.

Start Date * 25 25%

End Date * 25 25%

Chief Scientist * 25 25%

Vessel Code * 15 15%

Vessel Name * 5 5%

Community URI * 5 5%

Total: 100 %

Calculate the total score

Weigh the Cruise ID score with the combined cruise property score. The total of all weights should equal 100%.

Cruise ID Score Weight * 25 25%

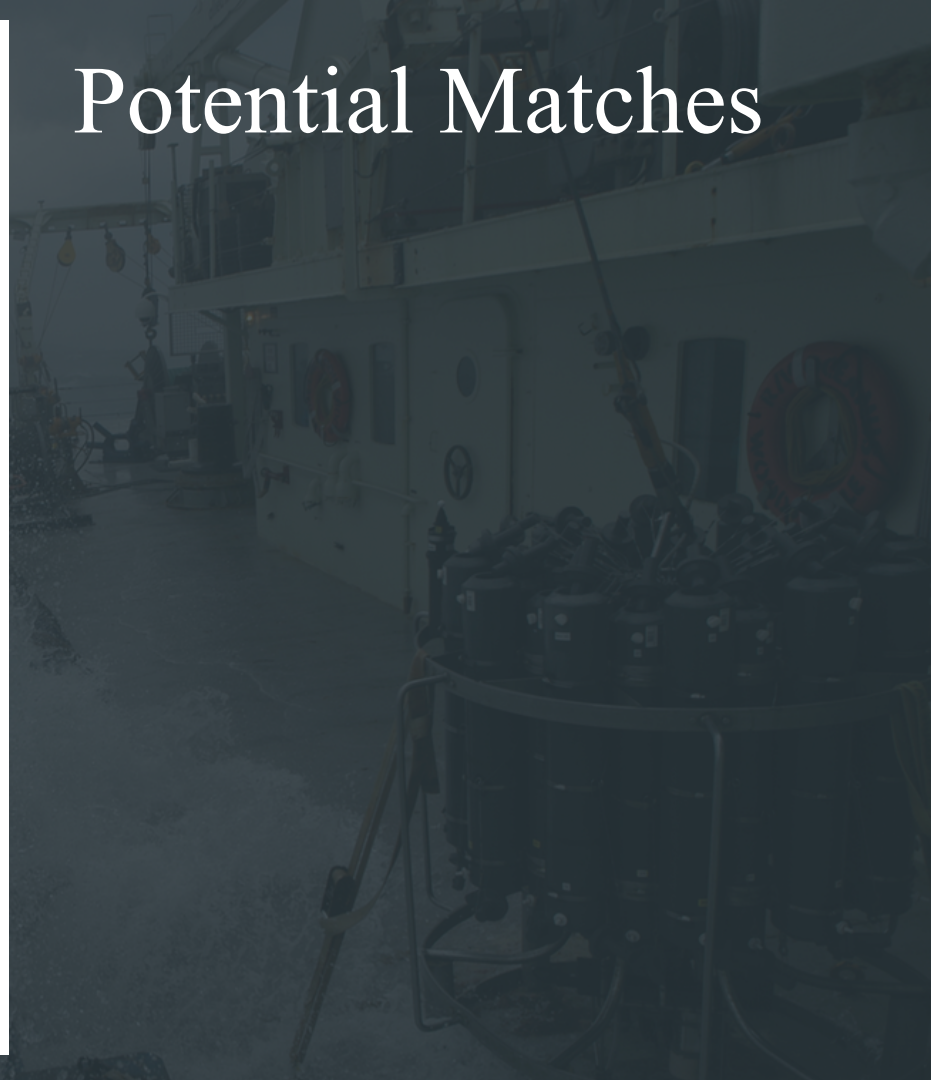
Combined Property Score Weight * 75 75%

Total: 100 %

Potential Matches: 26 of 491

Cruise ▲	Matches	Best Score	Existing Match
AT11-20	13	87%	
BH05-05	1	40%	
BH05-13	1	40%	
BH05-18	1	40%	
BH13-13	17	72%	
CB1007	1	81%	
CB1009	1	81%	
FK010	1	93%	
HRS1324	3	70%	
KM1416	1	100%	
KOK1115	4	77%	
LMG0102	19	100%	
LMG0414	6	77%	
LMG1312	5	100%	
MGL1216	11	96%	
NBB10-02	1		

Potential Matches



Potential Matches: 26 of 491

Cruise ^	Matches	Best Score	Existing Match
AT11-20	13	87%	
BH05-05	1	40%	
BH05-13	1	40%	
BH05-18	1	40%	
BH13-13	17	72%	
CB1007	1	81%	
CB1009	1	81%	
FK010	1	93%	
HRS1324	3	70%	
KM1416	1	100%	
KOK1115	4	77%	
LMG0102	19	100%	
LMG0414	6	77%	
LMG1312	5	100%	
MGL1216	11	96%	
NBB10-02	1		

Potential Matches

LMG0102 - Match to R2R Cruise

[^ back to results listing ^](#)

[<< Previous \(KOK1205\)](#)

[Next \(LMG0103\) >>](#)

BCO-DMO Cruise: LMG0102

LMG0102	score: 100.00%	
LMG0105	score: 74.07%	
LMG0106	score: 74.07%	
LMG0002	score: 73.93%	
LMG0802	score: 73.30%	
LMG0902	score: 73.30%	
LMG0104	score: 72.86%	
LMG0502	score: 72.32%	
LMG1102	score: 72.32%	
LMG0103	score: 72.25%	
LMG0602	score: 72.05%	
LMG0402	score: 71.88%	
LMG0109	score: 70.32%	
LMG0101	score: 70.18%	
LMG0107	score: 70.18%	
LMG0302	score: 68.63%	
LMG0108	score: 68.30%	
LMG0702	score: 66.43%	
LMG0202	score: 64.55%	

Potential Matches: 26 of 491

Cruise ^	Matches	Best Score	Existing Match
AT11-20	13	87%	
BH05-05	1	40%	
BH05-13	1	40%	
BH05-18	1	40%	
BH13-13	17	72%	
CB1007	1	81%	
CB1009	1	81%	
FK010	1	93%	
HRS1324	3	70%	
KM1416	1	100%	
KOK1115	4	77%	
LMG0102	19	100%	
LMG0414	6	77%	
LMG1312	5	100%	
MGL1216	11	96%	
NBB10-02	1		

Potential Matches

LMG0102 - Match to R2R Cruise

<< Previous (KOK1205)

^ back to results listing ^

BCO-DMO Cruise: LMG0102

LMG0102	score: 100.00%	100%
LMG0105	score: 74.07%	74%
LMG0106	score: 74.07%	74%
LMG0002	score: 73.93%	74%
LMG0802	score: 73.30%	73%
LMG0902	score: 73.30%	73%
LMG0104	score: 72.86%	73%
LMG0502	score: 72.32%	72%
LMG1102	score: 72.32%	72%
LMG0103	score: 72.25%	72%
LMG0602	score: 72.05%	72%
LMG0402	score: 71.88%	72%
LMG0109	score: 70.32%	70%
LMG0101	score: 70.18%	70%
LMG0107	score: 70.18%	70%
LMG0302	score: 68.63%	69%
LMG0108	score: 68.30%	68%
LMG0702	score: 66.43%	66%
LMG0202	score: 64.55%	65%

LMG0102 - Match to R2R Cruise

<< Previous (KOK1205)

^ back to results listing ^

Next (LMG0103) >>

BCO-DMO Cruise: LMG0102

LMG0102 score: 100.00%

Create Match

Match Type *

- Select a value -

Match Notes

Match Differences

Create Match

R2R Cruise Identifier: <http://data.rvdata.us/id/cruise/LMG0102>

BCO-DMO Cruise Identifier: [http://lod.bco-dmo.org/id/deployment/671723\(view in OSPREY\)](http://lod.bco-dmo.org/id/deployment/671723(view in OSPREY))

Property	BCO-DMO	R2R	Score
Cruise ID	LMG0102	LMG0102	100.00%
Start Date	2001-02-20	2001-02-20	100.00%
End Date	2001-03-14	2001-03-14	100.00%
Chief Scientist	Smith, Craig	Smith, Craig	100.00%
Vessel Name	Laurence M. Gould	Laurence M. Gould	100.00%
Vessel Code	33LG	33LG	100.00%
Community URI	http://vocab.nerc.ac.uk/collection/C17/current/33LG/	http://vocab.nerc.ac.uk/collection/C17/current/33LG/	100.00%

PROV for Comparing Matches over Time

AE1314	R2R Cruise: AE1314		100.00%	owl:sameAs	delete
AE1319	R2R Cruise: AE1319		100.00%	owl:sameAs	delete
AE1322	AE1322: R2R Cruise: AE1322		100.00%	owl:sameAs	delete
AE1409	AE1409: R2R Cruise: AE1409	End Date	98.13%	odo:matches	delete
AE1410	AE1410: R2R Cruise: AE1410	Start Date, End Date	94.38%	odo:matches	delete
AE1505	AE1505: R2R Cruise: AE1505		100.00%	owl:sameAs	delete
AE1516	AE1516: R2R Cruise: AE1516		100.00%	owl:sameAs	delete
AT11-04	AT11-04: R2R Cruise: AT11-04	Chief Scientist	94.08%	odo:matches	delete
AT11-07	AT11-07: R2R Cruise: AT11-07		100.00%	owl:sameAs	delete
AT11-17	R2R Cruise: AT11-17		100.00%	owl:sameAs	delete
AT11-30	R2R Cruise: AT11-30	Start Date	98.13%	odo:matches	delete
AT15-06	AT15-06: R2R Cruise: AT15-06		98.83%	owl:sameAs	delete
AT15-25	R2R Cruise: AT15-25	Chief Scientist	85.94%	odo:matches	delete
AT15-35	AT15-35: R2R Cruise: AT15-35		100.00%	owl:sameAs	delete
AT15-38	AT15-38: R2R Cruise: AT15-38		100.00%	owl:sameAs	delete
AT15-40	R2R Cruise: AT15-40		100.00%	owl:sameAs	delete

PROV-O + EMO helps us keep track:


Who asserted the match,

When,

Algorithm, Thresholds & Scores,

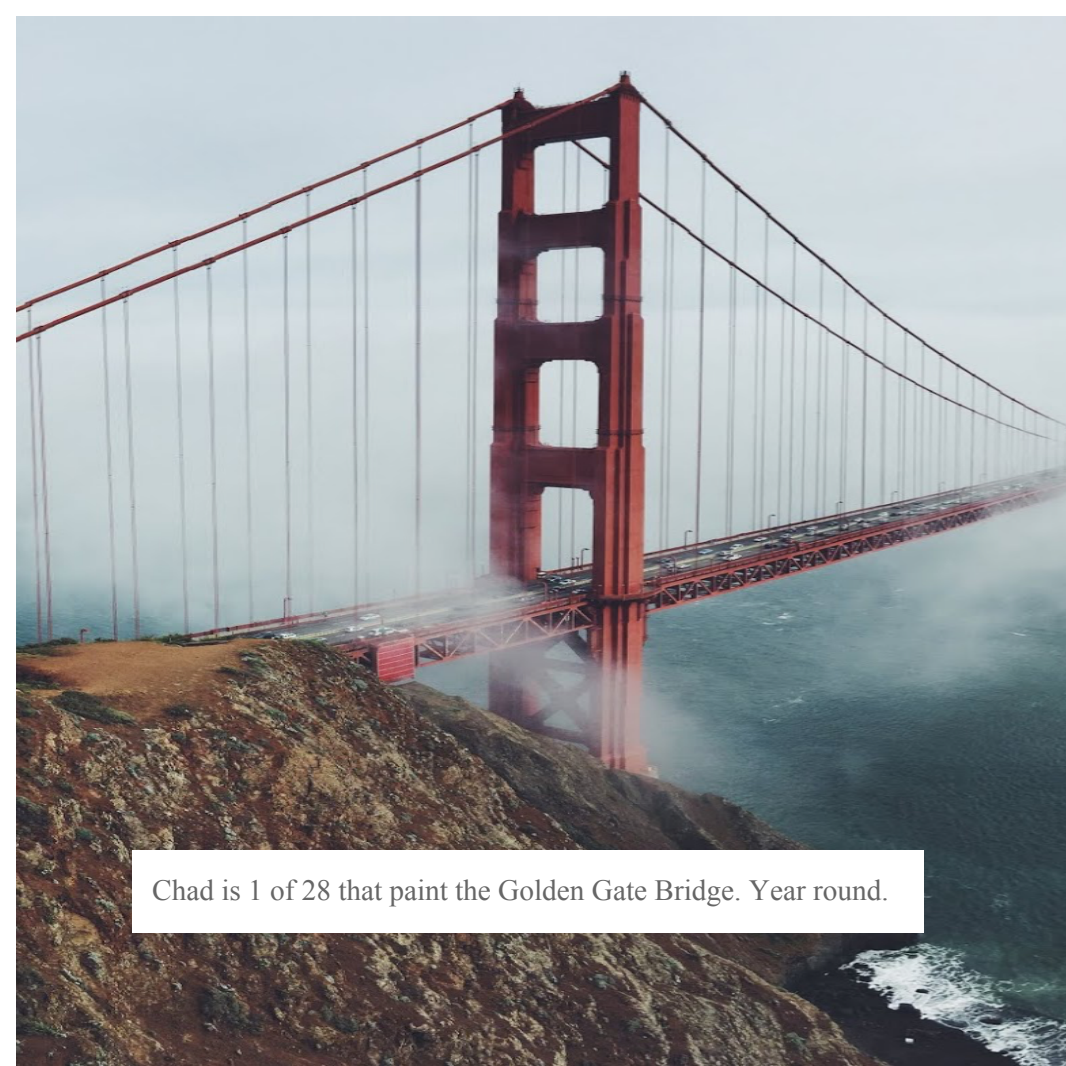
Original RDF,

and the EMO RDF.



"The job has its ups and downs. There's the swinging through the sky, the whale-watching, and the wielding of badass tools reminiscent of alien torture implements....[and] there's a weird kind of marine vertigo...."

- Chad Allan



Chad is 1 of 28 that paint the Golden Gate Bridge. Year round.

"The job has its ups and downs. There's the swinging through the sky, the whale-watching, and the wielding of badass tools reminiscent of alien torture implements....[and] there's a weird kind of marine vertigo...."

- Chad Allan

2,049

BCO-DMO Cruises.

6,555

R2R Cruises.

13,431,195

Comparisons.

28

Years.

15 seconds each, 40 hrs/wk, 50 wks/yr

28 years or 15 minutes...

Matching BCO-DMO Cruises to R2R Cruises



Time elapsed: 3 min 46 sec. Estimated time to finish: 11 min 41 sec

24%

Found potential matches for 128 BCO-DMO cruises

A dark, atmospheric photograph of a ship's deck, likely a research vessel, with the word "Questions?" overlaid in a white serif font. The ship's structure, including railings, equipment, and a red life preserver, is visible on the right side. The sea is choppy, and the sky is overcast. The overall tone is somber and contemplative.

Questions?